Video DNNs do not see motion illusions

Maechler, Marvin R (maechler@sas.upenn.edu) and **Soni, Ansh K (anshsoni@sas.upenn.edu)**
Department of Psychology, Goddard Labs
Philadelphia, Pennsylvania 19104 USA

Abstract

A key goal of computational neuroscience is to develop models that faithfully mimic human brain processing, including the efficiencies behind it. Visual illusions provide crucial test cases for this, as they are often the consequence of such efficiencies under biological constraints (e.g. efficient coding or optimal inference). Here we utilize illusions to investigate whether computer vision models process video inputs similarly to humans. Focusing on the double-drift illusion, we compare the representational geometry of these video models with behavioral and fMRI data from human subjects viewing the same stimuli. Representational similarity analyses reveal that while these models lack behavioral similarity to human observers, they do mimic the representational structure of some brain areas early in the visual processing hierarchy. Our findings demonstrate that, unlike humans, current vision models represent the physical stimulus at all points and do not combine motion and position information in a human-like manner. We thus find fundamental differences between human vision and DNNs in how temporal visual information is processed at later stages.

Keywords: computer vision; illusions; brain model similarity

Introduction

An accurate model of the visual system should represent stimuli in ways that parallel human visual processing, including reproducing the same systematic perceptual errors. Visual illusions provide an ideal test bed for evaluating visual models, as they can reveal the processing efficiencies employed by the brain (Eagleman, 2001). Furthermore, illusions allow researchers to dissociate physical stimuli from perceptual experience, enabling investigations into which representations (physical or perceived) are available to other cognitive processes such as attentional tracking or eye movements (Lisi & Cavanagh, 2015; Maechler, Cavanagh, & Tse, 2021; Maechler, Heller, Lisi, Cavanagh, & Tse, 2021).

This dissociation is particularly valuable because it permits matching stimuli in either their veridical, physical properties or their perceived properties, while creating differences in the alternative representational format. In early processing stages of the human visual system, the veridical stimulus is represented as it appears on the retina, while the illusory percept emerges in later stages (S. Liu, Yu, Peter, & Cavanagh, 2019; Li, Zeng, Shao, & Yu, 2023). If deep neural networks (DNNs) process visual information similarly to humans, we would expect them to represent stimuli according to their physical characteristics in early layers and according to their perceptual characteristics in later layers.

Previous research has established that certain convolutional neural network (CNN) architectures share encoding efficiencies with the human brain (Benjamin, Qiu, Zhang, Kording, & Stocker, 2019), and that these efficiencies can give rise to illusions such as the tilt illusion (Zhang, Mao, Aguirre, & Stocker, 2024). However, these findings are predominantly based on models processing static image inputs, whereas the human visual system must operate in a dynamic, continuously changing environment.

In this study, we examine video-processing DNNs to determine whether they ”experience” illusions. By comparing the representational structure of these models with human behavioral and fMRI data collected during stimulus presentation, we can assess whether current video DNNs capture fundamental aspects of human temporal visual processing.

Results

The representational geometry of DNN layer activity was more aligned with physical and not with perceptual stimulus representations for all DNNs we tested. The illusory video inputs were not represented like the perceptually matched (physically different) control stimuli. Generally, there was a moderate to high correlation between DNN layers and brain areas V1, V2, V3, V4 and MT, but not MST (Fig 1 C).

One crucial aspect of the double-drift illusion is its dependence on spatial uncertainty, as the illusion works only in the visual periphery (Kwon, Tadin, & Knill, 2015). Current DNNs incorporate neither the fovea-periphery trade-off present in human visual systems, nor do they combine position and motion information optimally when inferring trajectories of objects. Incorporating a simulated foveating step, following Freeman and Simoncelli (2011), where each video frame is blurred according to the distance from a simulated fixation cross 8 degrees of visual angle away from the illusion did not change the results.

Discussion

Our findings reveal a fundamental difference in how humans and current video DNNs process dynamic visual information. DNNs’ layerwise representations aligned with the physical rather than the perceived stimulus format. While these models shared representational similarities with early visual areas (V1-V4, MT), they failed to exhibit the perceptual integration that causes motion illusions in human observers.

Illusions perceived by the human visual system are often the consequence of neural efficiencies (Benjamin et al., 2019; Zhang et al., 2024) or of optimal inference using noisy inputs (Kwon et al., 2015). Accurate models of the visual system should benefit from the same computational efficiencies

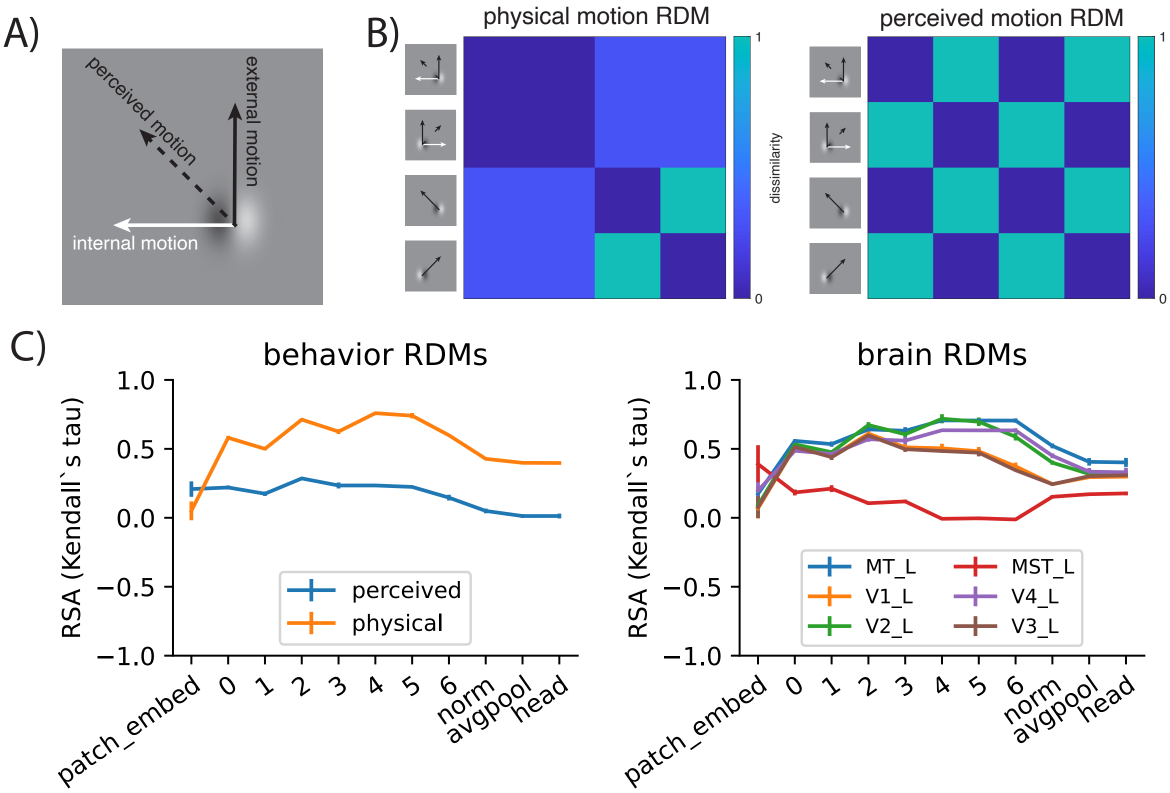

Figure 1: Method and Results A) Schematic of the illusion. External motion (translation of Gaussian envelope) and internal motion (Gabor phase drift) are mis-combined by the visual system to form an illusory perceived motion trajectory. B) Hypothetical RDMs for perceived and physical motion. Stimuli are either physically or perceptually matched. C) layerwise RSA results for behavioral and brain based RDMs from an example DNN (*SwinTransformer*). Brain data and stimuli are from Li et al. (2023).

present in biological vision, which would lead to similar systematic perceptual ”errors.” Current video models fail to reproduce these illusory percepts, suggesting they lack key computational principles that characterize human visual perception — particularly the integration of position and motion information under uncertainty.

Methods

Stimuli

The double-drift illusion arises when a Gabor patch has two conflicting motion signals: its envelope translating across the screen and its texture moving at a 90° angle internally (external and internal motion respectively). When viewed peripherally, observers perceive a trajectory that strongly deviates from the external motion path. They instead see the Gabor moving in a direction that is a weighted average of internal and external motion (Heller, Patel, Faustin, Cavanagh, & Tse, 2021). Computational modeling suggests this illusion results from optimal inference based on noisy motion and position estimates (Kwon et al., 2015). Here we used stimuli that closely matched those in experiment 1 from Li et al. (2023).

More motion illusions, such as the Flash-Lag illusion (Nijhawan, 1994) and motion-induced position shifts (De Valois & De Valois, 1991), will be included on the poster.

Video DNNs

In figure 1 we showcase a video *SwinTransformer* (Z. Liu et al., 2022, 2021) that does self-attention in multiple local windows as opposed to globally. The model uses the swin_tiny architecture and is trained on the Kinetics400 dataset (Kay et al., 2017), optimized for human action classification. The weights are taken from torchvision's default (TorchVision, maintainers, & contributors, 2016). More models (classification / next frame prediction / CNN / transformer) will be included on the poster.

Measures

We assessed the geometric nature of the representations generated by DNNs and brains with RSA. We use RSA as it allows us to compare how stimuli within a system are represented in relation to each other, both behaviorally and internally and see if that is consistent between models. While we cannot claim mechanistic similarity with this analysis (as similar representations do not require similar mechanisms), we can claim an inconsistency as similar representations are required for similar mechanisms.

Acknowledgments

We wish to thank Qing Yu for generously making the fMRI data from Li et al. (2023) available and for ensuring that the stimuli were as similar as possible to those used in their experiment. AKS was supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-2236662.

References

Benjamin, A., Qiu, C., Zhang, L.-Q., Kording, K., & Stocker, A. (2019). Shared visual illusions between humans and artificial neural networks. In *2019 conference on cognitive computational neuroscience.*

De Valois, R. L., & De Valois, K. K. (1991). Vernier acuity with stationary moving gabors. *Vision research*, *31*(9), 1619–1626.

Eagleman, D. M. (2001). Visual illusions and neurobiology. *Nature Reviews Neuroscience*, *2*(12), 920–926.

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature neuroscience*, *14*(9), 1195–1201.

Heller, N. H., Patel, N., Faustin, V. M., Cavanagh, P., & Tse, P. U. (2021). Effects of internal and external velocity on the perceived direction of the double-drift illusion. *Journal of Vision*, *21*(8), 2–2.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., . . . others (2017). The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Kwon, O.-S., Tadin, D., & Knill, D. C. (2015). Unifying account of visual motion and position perception. *Proceedings of the National Academy of Sciences*, *112*(26), 8142–8147.

Li, S., Zeng, X., Shao, Z., & Yu, Q. (2023). Neural representations in visual and parietal cortex differentiate between imagined, perceived, and illusory experiences. *Journal of Neuroscience*, *43*(38), 6508–6524.

Lisi, M., & Cavanagh, P. (2015). Dissociation between the perceptual and saccadic localization of moving objects. *Current Biology*, *25*(19), 2535–2540.

Liu, S., Yu, Q., Peter, U. T., & Cavanagh, P. (2019). Neural correlates of the conscious perception of visual location lie outside visual cortex. *Current Biology*, *29*(23), 4036–4044.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., . . . Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 10012–10022).

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., & Hu, H. (2022). Video swin transformer. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 3202–3211).

Maechler, M. R., Cavanagh, P., & Tse, P. U. (2021). Attentional tracking takes place over perceived rather than veridical positions. *Attention, Perception, & Psychophysics*, *83*(4), 1455–1462.

Maechler, M. R., Heller, N. H., Lisi, M., Cavanagh, P., & Tse, P. U. (2021). Smooth pursuit operates over perceived not physical positions of the double-drift stimulus. *Journal of Vision*, *21*(11), 6–6.

Nijhawan, R. (1994). Motion extrapolation in catching. *Nature*.

TorchVision, maintainers, & contributors. (2016). *Torchvision: Pytorch's computer vision library.* `https://github.com/pytorch/vision`. GitHub.

Zhang, L.-Q., Mao, J., Aguirre, G. K., & Stocker, A. A. (2024). The tilt illusion arises from an efficient reallocation of neural coding resources at the contextual boundary. *bioRxiv*.